On Topic Aware Recommendation to Increase Popularity in Microblogging Services (Short Paper)

Iouliana Litou[1(✉)], Vana Kalogeraki[1], and Dimitrios Gunopulos[2]

[1] Athens University of Economics and Business, Athens, Greece
{litou,vana}@aueb.gr
[2] University of Athens, Athens, Greece
dg@di.uoa.gr

Abstract. The flourish of Web-based Online Social Networks (OSNs) has led to numerous applications that exploit social relationships to boost the influence of content in the network. However, existing approaches focus on the social ties and ignore how the topic of a post and its structure relate to its popularity. Our work assists in filling this gap. The contribution of this work is two-fold: (i) we develop a scheme that automatically identifies the topic of a post, specifically tweets, in real-time without human participation in the process, and then (ii) based on the topic of the tweet and prior related posts, we recommend appropriate structural properties to increase the popularity of the particular tweet. By exploiting Wikipedia, our model requires no training or expensive feature engineering for the classification of tweets to topics.

1 Introduction

Content generation and propagation in Online Social Networks (OSNs) have flourished in recent years. The topic of content published in OSNs is inevitably linked to its diffusion and therefore identifying the topic of user generated content in OSNs may assist in a variety of applications, e.g., identifying influential users on a specific topic [1,6], personalizing web searches and recommendations [14,16] and discovering experts on the topic [12,13] to name a few. In this work we focus on the network of Twitter, one of the most popular social networks today, and aim to discover attributes that impact the popularity of the tweets in conjunction to their respective topic. We denote *popularity* as the number of retweets received by a tweet (also referred as *retweetability*). Extracting the topic of the tweets is particularly challenging as traditional text-classification methods for topic discovery are inappropriate for tweets due to the sparseness and non-standardization of the text. Moreover, tweets are concise texts that usually present misspellings, non-standard terms and noise. Challenges are further presented in terms of predicting the popularity of a tweet and making the appropriate recommendations, even if the topic is known, particularly since

C. Debruyne et al. (Eds.): OTM 2016 Conferences, LNCS 10033, pp. 673–681, 2016.
DOI: 10.1007/978-3-319-48472-3_40

not all factors contributing to the dissemination may be captured (e.g. external events).

We argue that current schemes proposed for discovering attributes that affect the popularity of the content in OSNs are inadequate, as they either ignore the topic or consider it as given. Contrary to the majority of the works in the literature that focus on either topic detection or popularity estimation in OSNs, we investigate these two aspects in concert. Our goal is to design a topic-aware recommendation system for micro-blogging services. We approach the problem in two phases: (i) initially, we identify topics related to a tweet by exploiting Wikipedia, and (ii) then we recommend the appropriate attributes to increase its retweetability. A topic aware recommendation can help deal with the information overload in micro-blogging communities, as information search may be filtered based on the topic. Additionally, by capturing the attributes that enhance the popularity of a tweet, under the context of the topic, we can contribute to its further propagation. We note that although we use Twitter, our approach is generic enough and can be exploited in several OSNs e.g., Facebook[1]. To the best of our knowledge, we are the first to consider both aspects of real-time tweet topic discovery and popularity in concert.

The contributions of our work are summarized as follows: (a) We propose a novel approach that exploits Wikipedia to automatically identify the topic(s) in the tweets in real-time, without the need for human agents or a prior knowledge base. Our scheme correctly identifies the topic of more than 89 % of the tweets, achieving an accuracy of up to 43 % higher than existing approaches. (b) We design a topic-based popularity prediction mechanism to estimate the probability of a tweet exceeding a popularity threshold and recommend appropriate structures to boost its popularity. (c) We evaluate the topic-aware recommendation mechanism on 76150 tweets from popular accounts and 85120 random tweets. Our experimental results illustrate the effectiveness of the proposed scheme, which correctly estimates tweet popularity with up to 91 % accuracy.

2 Our Approach

Our approach for solving the problem of topic-aware recommendations to increase popularity of a tweet unfolds in two phases. We initially deploy a classification mechanism to discover the set of topics covered in the tweet and later discover the features that impact the retweets from past tweets on the topics. The implementation of our Topic-Aware Recommendation System consists of two main components, (i) a Classification Component and (ii) a Recommendation Component. The overall system architecture is depicted in Fig. 1.

2.1 Classification Component

For the purposes of topic discovery we exploit Wikipedia[2] as a Knowledge Base. Wikipedia is a multilingual, web-based, free-content encyclopedia project

[1] https://www.facebook.com.
[2] https://en.wikipedia.org/wiki/Main_Page.

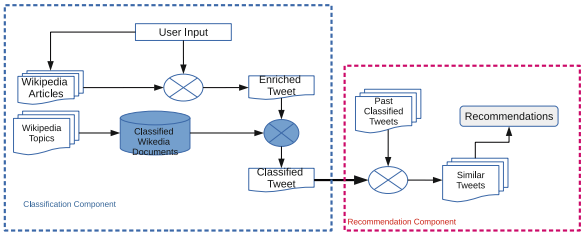


Fig. 1. System architecture

containing articles under various levels of concepts called categories. For each topic $c \in C$ in Table 1, we retrieve the set of Wikipedia articles D_c under this topic. We consider that these topics are generic enough, however more topics could be explored. For each article $d \in D_c$ we construct labelled document vectors. The label of the document vector denotes the corresponding topic of the article.

Table 1. Wikipedia classification main topics

Agriculture	Arts	Business	Chronology	Concepts
Culture	Education	Environment	Geography	Health
History	Humanities	Humans	Language	Law
Life	Mathematics	Medicine	Nature	People
Politics	Science	Society	Sports	Technology

Tweet Text Enrichment. Given a tweet t and the set of words $W = \{w_1, ..., w_n\}$ included in the text of t, $\forall w_i \in t$, excluding stop-words, we query the related Wikipedia articles using the Wikipedia API[3]. We consider consecutive words of the tweet starting with capitals both isolated and also as a single entity when querying Wikipedia. For each $w_i \in W$ the set of relevant articles A' are retrieved. A document vector V is later constructed for the tweet containing the words in W; this is augmented with the words included in each article $a' \in A'$. Using the *Enriched* tweet, we later deploy a classification algorithm to identify the topics C' of the tweet. Note, that a variety of classification algorithms may be exploited for the task. In all cases, the enrichment of the tweet significantly enhances performance, as we show in the experimental evaluation section. It is possible that a tweet may cover multiple topics, therefore we assume it may be related to at most k topics, where k is a tunable parameter. In our experiments we set $k = 3$, as a single topic might be misleading. On the other hand, more than 3 topics may lead to unrelated topics being identified.

[3] https://www.mediawiki.org/wiki/API:Main_page.

Classification Algorithm. A single document vector V_c for each topic c is computed by aggregating the document vectors of the articles on the topic. The similarity of the tweet to the topic is estimated as follows:

$$f(t,c) = \frac{1+|W_c|}{|W|} * \sum_{w \in W} sim(w, V_c) \tag{1}$$

where $\sum_{w \in W} sim(w, V_c)$ is the sum of similarities of words $w \in W$ in the tweet t classified to topic c, $|W_c|$ is the total number of words $w \in W$ classified to topic c and $|W|$ is the total number of words in tweet t. The similarity of the word w to topic c is computed as:

$$sim(w, V_c) = \frac{(\sum_{y}\{y : y \in w' \wedge y \in V_c\})^2}{|D_c|} \tag{2}$$

where y is a word in the augmented document vector w' of word w with the corresponding Wikipedia article and $|D_c|$ denotes the number of documents of topic c. The word w is classified in topic $c \in C$ so that $sim(w,c) = max\{sim(w, V_c) : \forall c \in C\}$. The above metric intuitively computes the common words of the augmented vector w' to the total words of a topic V_c, while taking into account the number of documents contained in c. In the experiments we also exploit the cosine similarity metric.

2.2 Recommendation Component

Given the classified tweet, the next step is to identify appropriate structures that positively affect popularity. This task is implemented through the recommendation component. The recommendation procedure includes the following steps: (i) initially the set of past similar tweets S, i.e., tweets covering at least one of the topics c' covered in the tweet are retrieved and a popularity threshold θ is estimated based on tweets in S, and (ii) appropriate structures are recommended to enhance retweetability as these are derived from the popular tweets in S.

The popularity threshold θ is estimated as $\theta = \frac{\bar{x}+\mu_{\frac{1}{2}}}{2}$ where $\bar{x}$ and $\mu_{\frac{1}{2}}$ are respectively the average and median value of retweets in S. We consider both values, as the average value may be affected by extreme values and most of the tweets are not frequently retweeted resulting in low median values. After θ is estimated, popular tweets in S, i.e. tweets with over θ retweets, are identified and a recommendation mechanism is triggered. To identify popular features and extract recommendations, we investigate different methodologies.

Conditional Probabilities (CP): Under CP the probability $P(p > \theta|a)$ of an attribute a being presented in popular tweets is estimated as:

$$P(p > \theta|a) = \frac{P(p \geq \theta \cap a)}{P(p \geq \theta)} = \frac{\sum_i t_i}{\sum_i pt_i} \tag{3}$$

where p is the number of retweets, $t_i = 1$ if tweet i in S are presents retweets above θ and attribute a is present, or $t_i = 0$ otherwise. $pt_i = 1$ if retweets of i exceed θ or $pt_i = 0$ otherwise. If $P(p > \theta|a) \geq 0.5$, then the attribute a is recommended. The Conditional Probabilities intuitively suggest the use of an attribute if it is present in popular tweets with high probability.

Decision Trees (DT): As CP assumes independence of attributes, an assumption that may not be necessarily valid, we further investigate the use of Decision Trees in recommendations. To construct the decision tree, the information gain of attribute a is estimated at each leaf as $IG(p \geq \theta, a) = H(p \geq \theta) - H(p \geq \theta|a)$, where $H(p \geq \theta)$ is the entropy of retweets exceeding θ and $H(p \geq \theta|a)$ is the conditional entropy of retweets given that the attribute a is present. Initially, we compute the $IG(p \geq \theta, a)$ for all attributes and for all tweets in S. Afterwards, the attribute a presenting the highest $IG(p \geq \theta, a)$ value is selected and tweets are split to S_a and S'_a, where S_a contains tweets in S presenting the attribute a and S'_a contains tweets in $S \setminus S_a$. The process is repeated for all attributes and a binary tree is constructed, with branches of the tree leading to leafs based on the presence or absence of attributes. The attributes leading to the leaf with the greatest proportion of popular tweets are recommended.

3 Experimental Evaluation

Our experimental evaluation focuses on estimating the performance of both the classification mechanism and the recommendation mechanism. We evaluate our approach using two real-world datasets, (a) a dataset of 76150 tweets from 20 **Popular** accounts[4] and (b) a dataset of 85120 **Random** tweets published in the UK area in January 2014 [18].

Classification Evaluation. In the first set of experiments we evaluate our classification scheme, referred as *Enriched* and compare it to the Support Vector Machine (SVM) classification model and a classifier that leverages Wikipedia structure (Wiki-Structure) to identify tweet topics [11]. We further conducted experiments by exploiting the cosine similarity measure between the enriched tweet and the documents $d_c \in D$, referred as Enriched-Cosine.

To verify our automated topic detection scheme we randomly select 1000 classified tweets from each dataset and asked users of CrowdFlower[5], a crowdsourcing tool, to select the topics relevant to the tweet according to the topics of Table 1. Each tweet is evaluated by 5 users. Tweets that users could not identify the topic are considered as conflicted. The results of the crowdsourcing evaluation are summarized in Table 2 and suggest that our automated topic-detection scheme correctly identifies the topic of the tweets on more than 69 % of all cases for the popular and 43 % of the random tweets. Excluding the cases that even human agents are unable to decide on the topic, i.e. conflicting tweets, our approach correctly identifies the topic in 89 % of the tweets from popular accounts

[4] Best People On Twitter: http://goo.gl/AOU0GU.
[5] http://www.crowdflower.com/.

Table 2. Classification evaluation

	Popular	Random
Correct	692	437
Incorrect	80	24
Conflicted	228	539

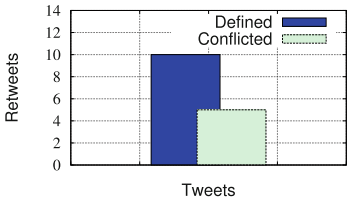
Table 3. Comparison of classification methods

	Popular		Random	
	Correct	Incorrect	Correct	Incorrect
Enriched	692	80	437	24
SVM	321	451	134	327
Wiki-Structure	129	643	53	408
Enriched-Cosine	354	418	239	222

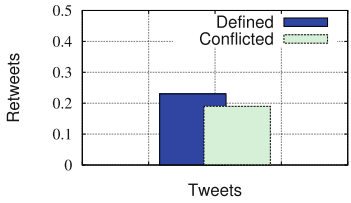
and over 94 % of the random tweets. As tweets for the evaluation are randomly selected, we note that results can be generalized for the entire datasets. In Table 3 we present the results of our suggested methodology compared to the rest of the classification techniques. As the results indicate, the enrichment of the tweets significantly enhances the performance of topic discovery. Moreover, considering the common terms to the topic rather than the term frequencies further assists the process, as suggester by the performance of our suggested metric over the cosine similarity. Our approach manages to outperform the rest by up to 43 % in accurately identifying the topic of a tweet for the popular accounts and over 42 % in the random tweets case.

To show how the clarity of the topic impacts on the popularity of the tweet, we present in Fig. 2a the median value of retweets received by tweets from popular accounts. The median is estimated separately for tweets with conflicted and tweets with clearly defined topic, based on the results of the CrowdFlower evaluation. The results suggest that tweets with clearly defined topic are more popular. Furthermore, tweets from popular accounts seem to have overall more defined topic compared to random tweets (Table 2), implying that the clearly defined topic of the tweets may have assisted in the popularity of the accounts. We note that the median value for the random tweets equals zero, therefore in Fig. 2b we present the average value of retweets.

Topic-Based Recommendation Evaluation. The last set of experiments focuses on estimating the performance of the recommendation mechanisms. We used 20 % of the tweets of each topic for evaluation and 80 % of the tweets to extract recommendations. The results are presented in Table 4. As the datasets are imbalanced, including few popular tweets, we provide the F1-measure and the Balanced Accuracy. Conditional probabilities lead to high precision, yet many popular tweets are considered as unpopular as suggested by the Recall value.



(a) Popular account retweets median



(b) Random Tweets retweets average (median equals 0)

Fig. 2. Retweets based on topic clarity

Table 4. Result on recommendation mechanisms evaluation

Dataset	Recommendation	Precision	Recall	Accuracy	F1 Measure	Balanced accuracy
Technology (Popular)	Random	0.4769	0.0650	0.5275	0.1144	0.5010
	Conditional	0.6817	0.3480	0.6176	0.4608	0.6021
	Decision tree	0.6939	0.8661	0.7441	0.7705	0.7450
Education (Popular)	Random	0.5926	0.0627	0.5229	0.1135	0.5109
	Conditional	0.9068	0.4196	0.6966	0.5737	0.6894
	Decision tree	0.8521	0.8327	0.8435	0.8423	0.8436
Geography (Random)	Random	0.2083	0.02538	0.8871	0.0452	0.5070
	Conditional	0.1216	0.4112	0.6249	0.1877	0.5313
	Decision tree	0.0331	0.0976	0.9176	0.0494	0.5168

The performance of Decision Trees over Conditional Probabilities implies that attributes may not be independent. For the popular accounts dataset recommendations based on the Decision Trees correctly identify attributes that lead to popularity of tweets with accuracy of over 74 % on the topic of Technology and over 84 % on the topic of Education. As most of the tweets are unpopular in the random dataset, few tweets are correctly identified as popular based on the suggestions. Yet, the accuracy is up to 91 %.

4 Related Work

In [4] the authors develop a method for analysing and automatically classifying publications using the Wikipedia category hierarchy. In [5] Wikipedia is queried based on entities of the tweets and the snippet of the retrieved articles is added to the text of the tweet to cluster tweets. Their goal is to identify trend-topics of the clusters. In [11] Wikipedia is used to disambiguate and categorize the entities of a tweet and to develop a "topic profile" that characterizes users' interest. Genc *et al.* in [7] extend this work to calculate between-tweet distances by exploiting the Wikipedia structure. In [3] Baralis *et al.* suggest a multi-level classification technique for clustering tweets using the Vector Space Model and association

rules between terms of the documents. Banerjee *et al.* in [2] cluster news feeds using Wikipedia to enrich the text and improve clustering. In [10] the authors demonstrate that a reasonably effective classifier can be created to identify the informative nature of tweets based on crowdsourcing data. In [21] authors present a method to classify Twitter user interests using time series generated from the contents of tweet streams. Towards identification of user interests and latent topics in document collections authors in [8] propose feature and social based topic models. Twitter profiles are also inferred using LDA topic model in [15]. Shu *et al.* in [17] aim at discovering the content and contextual features that impact retweetability. Zaman *et al.* in [22] develop a probabilistic model using a Bayesian approach to predict the diffusion of a tweet and the time lapse needed to be propagated. In [20] authors identify properties of the tweets that lead to predictions of information propagation. Wang *et al.* in [19] propose a recommendation system for expanding the propagation of a tweet through mentions. Kiciman in [9] studies the factors that influence self-reporting bias on twitter and found that the extremeness of an event accounts to tweet rate. However, none of the above works consider the topic of the tweets in association to their popularity. Aslay *et al.* in [1] consider topics to identify the most appropriate seeds for increasing propagation of the contents in OSNs, yet topic discovery is not addressed in their work.

5 Conclusions

In this paper we aim at understanding the factors that associate with retweeting under the context of the topic of the tweets and propose a topic-based recommendation system for increasing tweet popularity. We show that our approach performs with up to 68 % better accuracy compared to existing approaches in identifying tweet topics and effectively estimates the tweet's popularity with up to 91 % accuracy.

Acknowledgments. This research has been financed by the European Union through the FP7 ERC IDEAS 308019 NGHCS project and the Horizon2020 688380 VaVeL project.

References

1. Aslay, Ç., Barbieri, N., Bonchi, F., Baeza-Yates, R.A.: Online topic-aware influence maximization queries. In: EDBT 2014, Athens, Greece, pp. 295–306, March 2014
2. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: SIGIR 2007, pp. 787–788. ACM (2007)
3. Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L., Xiao, X.: Analysis of twitter data using a multiple-level clustering strategy. In: Cuzzocrea, A., Maabout, S. (eds.) MEDI 2013. LNCS, vol. 8216, pp. 13–24. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41366-7_2
4. Biuk-Aghai, R., Ng, K.K.: A method for automated document classification using wikipedia-derived weighted keywords. In: 2014 International Conference on ICODSE, pp. 1–6, November 2014

5. Chen, Q., Shipper, T., Khan, L.: Tweets mining using wikipedia and impurity cluster measurement. In: 2010 IEEE ISI, pp. 141–143 (2010)
6. Chen, S., Fan, J., Li, G., Feng, J., Tan, K.-L., Tang, J.: Online topic-aware influence maximization. Proc. VLDB Endow. **8**(6), 666–677 (2015)
7. Genc, Y., Sakamoto, Y., Nickerson, J.V.: Discovering context: classifying tweets through a semantic transform based on wikipedia. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011. LNCS (LNAI), vol. 6780, pp. 484–492. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21852-1_55
8. Hu, B., Song, Z., Ester, M.: User features and social networks for topic modeling in online social media. In: Proceedings of the 2012 ASONAM, ASONAM 2012, pp. 202–209 (2012)
9. Kiciman, E.: OMG, i have to tweet that! a study of factors that influence tweet rates. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z. (eds.) ICWSM. The AAAI Press (2012)
10. Machedon, R., Rand, W., Joshi, Y.: Automatic crowdsourcing-based classification of marketing messaging on twitter. In: SocialCom 2013, pp. 975–978 (2013)
11. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data. ACM (2010)
12. Munger, T., Zhao, J.: Identifying influential users in on-line support forums using topical expertise and social network analysis. In: Proceedings of the 2015 IEEE/ACM ASONAM, pp. 721–728 (2015)
13. Niu, W., Liu, Z., Caverlee, J.: LExL: a learning approach for local expert discovery on twitter. In: Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Nunzio, G.M., Hauff, C., Silvello, G. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 803–809. Springer, Heidelberg (2016). doi:10.1007/978-3-319-30671-1_71
14. Ozsoy, M.G., Polat, F., Alhajj, R.: Modeling individuals and making recommendations using multiple social networks. In: Proceedings of the 2015 IEEE/ACM ASONAM, pp. 1184–1191 (2015)
15. Quercia, D., Askham, H., Crowcroft, J.: Tweetlda: supervised topic classification and link prediction in twitter. In: Proceedings of the 4th Annual ACM WebSci, WebSci 2012, pp. 247–250. ACM, New York (2012)
16. Shafiq, M.O., Alhajj, R., Rokne, J.G.: On personalizing web search using social network analysis. Inf. Sci. **314**, 55–76 (2015)
17. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: SOCIALCOM 2010, pp. 177–184. IEEE Computer Society (2010)
18. Valkanas, G., Gunopulos, D.: Location extraction from social networks with commodity software and online data. In ICDMW (2012)
19. Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W., Cai, D., He, X.: Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems. In: WWW 2013, pp. 1331–1340 (2013)
20. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in Twitter. In: 4th International AAAI ICWSM, May 2010
21. Yang, T., Lee, D., Yan, S.: Steeler nation, 12th man, and boo birds: classifying twitter user interests using time series. In: Proceedings of the 2013 IEEE/ACM ASONAM, pp. 684–691 (2013)
22. Zaman, T., Fox, E.B., Bradlow, E.T.: A bayesian approach for predicting the popularity of tweets. CoRR (2013)